STAT2

Xingyu Chen

February 28, 2022

Fitting value is anther name of predicted value.

Y-hat or $\widehat{y}$ means the predicted value of y in a regression equation, or the average value of the response variable.

Multiple linear regression (MLE): $\widehat{\varepsilon_i} = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{i,1} - ... - \widehat{\beta}_{(p-1)} X_{i,(p-1)}$

Consider the system of linear equations

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j} + \varepsilon_i$$

for $i = 1, \ldots, n$, where n is the number of data points (measurements in the sample), and $j = 1, \ldots, p$, where

1. $p + 1$ is the number of parameters in the model.
2. Y_i is the i^{th} measurement of the response variable.
3. $X_{i,j}$ is the i^{th} measurement of the j^{th} predictor variable.
4. ε_i is the i^{th} error term and is a random variable (often assumed to be $N(0, \sigma^2)$).
5. β_j are unknown parameters of the model, $(j = 0, \ldots, p)$. We hope to estimate these, which would help us characterize the relationship between the predictors and response.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \end{bmatrix} = \begin{bmatrix} x_{1,0} & x_{1,1} & \cdots & x_{1,j} \\ x_{2,0} & x_{2,1} & \ddots & x_{2,j} \\ & \vdots & & \vdots \\ x_{i,0} & x_{i,1} & \cdots & x_{i,j} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_j \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \end{bmatrix}$$

The pdf of normal distribution is:

$$N(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The MLE for μ^2 will be:

$$L(\mu^2 \mid \theta) = \prod_i^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = (\frac{1}{\sqrt{2\pi\sigma^2}})^n \prod_i^n e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Take the log for both side:

$$\ln L(\mu^2 \mid \theta) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{\sum(x-\mu)^2}{2\sigma^2}$$

The take the derivative for both side:

$$\frac{\partial \ln L(\mu^2 \mid \theta)}{\partial \mu^2} = \frac{1}{\sigma^2}\sum_i^n (x_i - \mu) \equiv 0$$

Thus,

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_i^n x_i$$

we have:

$$\bar{x} = \sum_i^n \frac{x_i}{n}$$

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta = E(\bar{x}^2) - \mu^2 = Var[\bar{x}] + E[\bar{x}]^2 - \mu^2$$

$$= \frac{\sigma^2}{n} + \mu^2 - \mu^2 = \frac{\sigma^2}{n}$$

Thus,

$$\text{Bias}(\hat{\theta}) = \frac{\sigma^2}{n}$$

$$\text{Var}(\hat{\theta}) = \text{Var}(\mu^2) = \text{Var}(\bar{x}^2)$$

The MGF of $\bar{x}$ is $e^{\mu t + \frac{t^2\sigma^2}{2n}}$.
Thus,

$$m'(t) = (\mu + \frac{t\sigma^2}{n}) * e^{\mu t + \frac{t^2\sigma^2}{2n}}$$

$$m''(t) = (\mu + \frac{t\sigma^2}{n})^2 * e^{\mu t + \frac{t^2\sigma^2}{2n}} + \frac{\sigma^2}{n} * e^{\mu t + \frac{t^2\sigma^2}{2n}}$$

$$m'''(t) = (\mu + \frac{t\sigma^2}{n})^3 * e^{\mu t + \frac{t^2\sigma^2}{2n}} + \frac{3\sigma^2(\mu + \frac{t\sigma^2}{n})}{n} * e^{\mu t + \frac{t^2\sigma^2}{2n}}$$

$$m''''(t) = (\mu + \frac{t\sigma^2}{n})^4 * e^{\mu t + \frac{t^2\sigma^2}{2n}} + \frac{6\sigma^2(\mu + \frac{t\sigma^2}{n})^2}{n} * e^{\mu t + \frac{t^2\sigma^2}{2n}} + \frac{3\sigma^4}{n^2} * e^{\mu t + \frac{t^2\sigma^2}{2n}}$$

$$\text{Var}(\hat{\theta}) = m''''(0) - [m''(0)]^2 = \mu^4 + \frac{6\sigma^2\mu^2}{n} + \frac{3\sigma^4}{n^2} - (\mu^2 + \frac{\sigma^2}{n})^2 = \frac{4\sigma^2\mu^2}{n} + \frac{2\sigma^4}{n^2}$$

Bootstrap:

Pros: It is a straightforward way to derive estimates of confidence intervals for complex estimators of the distribution. It is also an appropriate way to control and check the stability of the results. It is asymptotically more accurate than the standard intervals obtained using sample variance and assumptions of normality.

Cons: Bootstrapping depends heavily on the estimator used and, though simple, ignorant use of bootstrapping will not always yield asymptotically valid results and can lead to inconsistency. The result may depend on the representative sample.

χ^2 confidence interval is used when the data are normal, not for gamma.

The interval that I can be 90% certain contains the population standard deviation.

False: For a fixed sample size $n = 20$, as B increases, $\widehat{\text{Bias}}(\hat{\theta})$ will approach $\text{Bias}(\hat{\theta})$. That is, for a fixed n, the bootstrap estimate of the bias will approach the true bias as the number of bootstrap samples, B increases.

Sample means normally distribute assumptions when we compute a 95% confidence interval for the population mean score.

The confidence interval from bootstrap is not valid because bootstrap is not dependent on normal values on not iid (independent and identically distributed).

Ordinary least square (OLS): $\hat{\beta} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$

The columns of X are linearly independent so the 'Gram' matrix X^TX to be invertible. Individuals (observations) are independent. Each features are independent each other.No column can be written as a linear combination of the others. X has full rank.

Suppose that the number of measurements (n) is less than the number of model parameters ($p + 1$). What does this say about the invertibility of X^TX ? What does this mean on a practical level? Some variables are removed from the model, either because they are constant or because they belong to a block of collinear variables. The theoretical limit is n-1, as with greater values the X'X matrix becomes non-invertible. Or in other words, the columns of X are not linearly independent.

What is true about about $\widehat{\beta}$ if $X^T X$ is not invertible? We can not find N different values of $\widehat{\beta}$. We know longer have a system of linearly independent equations because X is no longer full rank.

Gauss–Markov theorem: the ordinary least squares (OLS) estimator has the lowest sampling variance within the class of linear unbiased estimators, if the errors in the linear regression model are uncorrelated, have equal variances and expectation value of zero. The errors do not need to be normal, nor do they need to be independent and identically distributed (only uncorrelated with mean zero and homoscedastic with finite variance). The requirement that the estimator be unbiased cannot be dropped, since biased estimators exist with lower variance.

Proof of Gauss-markov theorem shows that $var(a^T y) \geq var(c^T \widehat{\beta})$. The equals part happens when $a^T y = \lambda^T X^T y$, which implies that the OLS estimator is not only **minimum variance** but is also **unique**.

Scale changes: xi -> xi+a /b, beta -> b * beta

When discuss **Collinearity** in a linear model of A against B, we defined an metric R_i^2: The R^2 value when you run a linear model of x_i against all other predictors.

X^TX is not singular but close. Correlation matrix: if R^2 = 1, then x can be accurately describe by other x. X^TX, lambda, eigenvalues, if any lambda = 0, collinear warning, if multiple lambda small then it is multicollinearity. var(beta) = sigma^2 (1/ 1-R$^{2)(1/(x_j-x_i)}$2), variance inflation factor R^2 = 1 -> VIF goes up.

$$H = X\left(X^T X\right)^{-1} X^T$$

The goal of this question is to use the hat matrix to prove that the fitted values, $\widehat{\mathbf{Y}}$, and the residuals, $\widehat{\varepsilon}$, are uncorrelated.

Show that $\widehat{Y} = HY$. That is, H "puts a hat on" Y. Using least square estimator

$$\widehat{\beta} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T Y$$

Using our equation for $\widehat{\beta}$, we then have

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\beta} = \mathbf{X}\left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{H}\mathbf{Y}$$

Show that H is symmetric: $H = H^T$. Multiplying $\mathbf{X}$ by $\mathbf{H}$, we obtain

$$\mathbf{H}\mathbf{X} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{X} = \mathbf{X}$$

Show that $H\left(I_n - H\right) = 0_n$, where 0_n is the zero matrix of size $n \times n$. The vector of residuals, $\mathbf{e}$, is

$$\mathbf{e} \equiv \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

the expected residual vector is zero:

$$\mathbb{E}[\mathbf{e}] = (\mathbf{I} - \mathbf{H})(\mathbf{X}\beta + \mathbb{E}[\epsilon]) = \mathbf{X}\beta - \mathbf{X}\beta = 0$$

Stating that $\widehat{\mathbf{Y}}$ is uncorrelated with $\hat{\varepsilon}$ is equivalent to showing that these vectors are orthogonal.* That is, we need to show that their dot product is zero:

$$\widehat{\mathbf{Y}}^T\hat{\varepsilon} = 0$$

Prove this result.

$$\sum \widehat{\mathbf{Y}}^T\hat{\varepsilon} = \sum \widehat{\mathbf{Y}}\hat{\varepsilon} = \sum(\widehat{\beta_0} + \widehat{\beta_1}X_i)\hat{\varepsilon} = \sum \widehat{\beta_0}\hat{\varepsilon} + \widehat{\beta_1}X_i\hat{\varepsilon} = 0 + 0 = 0$$

Why is this result important in the practical use of linear regression? Because $\widehat{\mathbf{Y}}^T\hat{\varepsilon} = 0$, the residuals satisfy $\text{rank}(X) = p + 1$ linear equalities. Hence, although there are n of them, they are effectively $n - p - 1$ of them. The number $n - p - 1$ is therefore referred to as the degrees of freedom of the residuals $\hat{e}_1, \ldots, \hat{e}_n$.

For SLR: MLE and OLS will give you the same estimator
For MLR: OLS will always give you a better estimator than MLE

An error is the difference between the observed value and the true value (very often unobserved, generated by the DGP).

A residual is the difference between the observed value and the predicted value (by the model).

Consider $\widehat{\beta}$, what is $\sigma^2_{\widehat{\beta}}$. $\sigma^2(X^TX)^{-1}$ in terms of dimensionality.

lmod <- lm(Species ~ Endemics + Elevation + Nearest + Adjacent, gala)
summary(lmod)

for every 1 unit increase in Species, the endemics goes up by 4.192551. Can vary by 0.429056.Pr(>|t|) represents the p-value associated with the value in the t value column. If the p-value is less than a certain significance level (e.g. 0.05) the the predictor variable is said to have a statistically significant relationship with the response variable in the model. Here is 5.1 * 10^-10 so Endemics have a statistically significant relationship with the Species in the model.

nullmod <- lm(Species ~ 1, gala)
anova(nullmod, lmod)

R^2 represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Approximately 94% of the observed variation can be explained by the model's inputs.

Plot the residuals vs the fitted values. They are uncorrelated. This plot is used to detect non-linearity, unequal error variances, and outliers.

1 Module 1.1 Linear Regression

A **statistical** unit is one member of the set of entities being studied

A **population** is a collection of units about which research questions are asked

A **sample** is a subset of the population. Typically, samples should be representative

Inferential statistics and data science is the process of learning about relationships in a sample in a way that is reliable enough to generalize from the sample to a population of interest.

To **operationalize a concept** means to derive a set of steps to measure the concept

The **validity** of a dataset or measurement tool is the extent to which the dataset or measurement tool measures what it claims to measure

linear regression: Is used to explain or model the relationship between a single variable Y, and one or more variables $X_1, \ldots, X_p$

- Y is called the response, outcome, output, or dependent variable
- $X_1, \ldots, X_p$ are called predictors, inputs, independent variables, explanatory variables, or features. In some contexts, they are also called covariates.

Regression analysis has two main objectives:

- **Prediction**: predict an unmeasured/unseen Y using observed $X_1, ..., X_p$
- **Explanation**: To assess the effect of, or explain the relationship between, Y and $X_1, ..., X_p$.

$\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ be the response variable and $\mathbf{x_1} = \begin{pmatrix} x_{1,1} \\ x_{2,1} \\ \vdots \\ x_{n,1} \end{pmatrix}, \mathbf{x_2} = \begin{pmatrix} x_{1,2} \\ x_{2,2} \\ \vdots \\ x_{n,2} \end{pmatrix}, \ldots, \mathbf{x_p} = \begin{pmatrix} x_{1,p} \\ x_{2,p} \\ \vdots \\ x_{n,p} \end{pmatrix}$ be predictors; we will collect the predictors in a matrix: $X = (1x_1x_2...x_p)$, where 1 = (1,1,. . . ,1)ˆT. Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ be a vector of parameters. Finally, let $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ be a vector of error terms.

Definition/Assumptions of the linear regression model:
1. **Linearity** xian xing. X ~ Y scatter plot follows a Linear pattern.
2. **Independence** du li xing. Y is independent of errors/residuals.
3. **Homoskedasticity (constant variance)** fang cha qi xing. variance is the same for all X.
4. **Normality** zheng tai xing. residuals approximately normally distributed, with a mean of zero.

Interpreting simple linear regression parameters:
1. β_0: the intercept of the true regression line. β_0 is the average value of Y when x is zero. Usually this is called the "baseline average".
2. β_1: the slope of the true regression line. β_1: is the average change in Y associated with a 1-unit increase in the value of x.

Interpreting multiple linear regression parameters:
$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i(p-1)} + \varepsilon$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon} \\ \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$F = kx$: F = Force; k = Spring Constant; x = Displacement

A **circular analysis or double dipping** is the process of exploring a dataset in an attempt to discover what relationships exist, and then test hypotheses related to that exploration on the same dataset.

Ways to avoid circular analyses:
1. Design the analysis and prespecify research hypotheses before observing the data.
2. Subset the data

2 Module 1.2 Least squares estimation

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

: y is measured response; B_0 and B_1, E_i are unknown, to be estimated; X_i is measured predictors

The **line of best fit** to the data is the line that minimizes the sum of the squared vertical distances between the line y and the observed points

$y = X\beta + \varepsilon$: The problem is to find a B so that XB is as close as possible to y.

The **surface of best** fit to the data is the surface that minimizes the sum of the squared vertical distances between the surface and the observed points:

Let X be an $m \times n$ matrix, $\mathbf{v}$ be $n \times 1$, and $\mathbf{y}$ be $m \times 1$. Then:
1. Lemma 1: Then $X^T X$ is symmetric, i.e., $\left(X^T X\right)^T = X^T X$.
2. Lemma 2: Let $\mathbf{y} = X\mathbf{v}$. Then $\frac{\partial y}{\partial \mathbf{v}} = X$ and $\frac{\partial y^T}{\partial \mathbf{v}} = X^T$
3. Lemma 3: Let $c = \mathbf{v}^T \left(X^T X\right) \mathbf{v}$. Then $\frac{\partial c}{\partial \mathbf{v}} = 2X^T X\mathbf{v}$

The **residuals** are defined as:

The **fitted values** are defined as:

The **hat matrix**, H, is defined as:

Least Squares Estimation: We define the best estimate of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ as the one that minimized the sum of the squared residuals:

In order to use least squares, we assume that: 1. $E(\varepsilon_i) = 0$ for all $i = 1, \ldots, n$.
2. $E(Y_i) = \mathbf{x_i}^T \boldsymbol{\beta}$ for all $i = 1, \ldots, n$.
3. $\text{Cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} 0 & i \neq j \\ \sigma^2 & i = j \end{cases}$
4. $\left(X^T X\right)^{-1}$ exists.

The **Gauss-Markov Theorem**: Suppose that:
1.E $(\varepsilon_i) = 0$ for all $i = 1, \ldots, n$.
2. $E(Y_i) = \mathbf{x_i}^T \boldsymbol{\beta}$ for all $i = 1, \ldots, n$.
3. $\text{Cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} 0 & i \neq j \\ \sigma^2 & i = j \end{cases}$
4. $\left(X^T X\right)^{-1}$ exists.
Then $\widehat{\boldsymbol{\beta}}$ is the "best" unbiased estimator of $\boldsymbol{\beta}$.

The **maximum likelihood estimator**.
Suppose that $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Then:
1. marginal pdf:
2. joint pdf:
3. log-likelihood:

Sums of squares:
1. RSS: Residual sum of squares:
2. ESS: Explained (or regression) sum of squares:
3. TSS: Total sum of squares:

The residual sum of squares **RSS** can be interpreted as a measure of how much variation in y is left unexplained by the model—that is, how much cannot be attributed to a linear relationship.

The parameter σ^2 determines the amount of spread about the true regression line.

An estimate of σ^2 will be used in statistical inference (e.g., confidence interval formulas and hypothesis testing), presented in the next two sections.

Note:
1. The divisor $n-(p+1)$ in is the number of degrees of freedom (df) associated with RSS and $\hat{\sigma}^2$.
2. The RSS has $n-(p+1)$ df because $p+1$ parameters must first be estimated to compute it, which results in a loss of $p+1$df.
3. Replacing each y_i in the formula for $\widehat{\sigma}^2$ by the r.v. Y_i gives a random variable.
4. It can be shown that the r.v. $\hat{\sigma}^2$ is an unbiased estimator for $\widehat{\sigma}^2$.

The coefficient of determination, R^2, is defined as:

$$R^2 = 1 - \frac{RSS}{TSS}$$

Note: - $0 \leq R^2 \leq 1$
- Assuming that the model is correct, R^2 is interpreted as the proportion of observed variation in y explained by the model.

Warnings about R^2
1. R^2 can be close to 1 but the model is the wrong fit for the data.
2. R^2 can be close to 0 even when the model is the correct fit for the data.
3. R^2 should not be used to compare models with a different number of predictors.
4. R^2 says nothing about the causal relationship between the predictors and the response.

The least squares estimate is the solution to the normal equations:

$$X^T X\boldsymbol{\beta} = X^T\mathbf{Y}.$$

1. When $\left(X^T X\right)^{-1}$ exists, there is a unique solution, $\widehat{\boldsymbol{\beta}}$.
2. When $\left(X^T X\right)^{-1}$ does not exist, there will be infinitely many solutions.

Definition: When $\left(X^T X\right)^{-1}$ does not exist, the regression model is said to be non-identifiable (or, unidentifiable).

Why might we have non-identifiability?
1. One variable is just a multiple of another.
2. One variable is a linear combination of several others.

3. There are more variables than members in the sample.
Note: Near non-identifiability is trickier than exact non-identifiability.

OLS (Ord. least Squares):
For SLR,

$$\varepsilon_i = y_i - E(y_i) = y_i - (\beta_0 + \beta_1 x_i)$$

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

which is equivalent to the matrix form

$$\mathcal{Q} = \varepsilon'\varepsilon = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

$$\frac{\partial Q}{\partial \beta_0} = 2\sum(y_i - \beta_0 - \beta_1 x_i)(-1) \equiv 0$$

Now that the β values that will minimize $\mathcal{Q}$ are computed, the fitted regression line is written

$$\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

where estimated (predicted) errors, also called residuals, are defined to be

$$\hat{\varepsilon}_i = Y_i - \widehat{Y}_i$$

Several properties of the fitted regression line will be helpful in understanding the relationships between $\mathbf{X}, \boldsymbol{\beta}, \varepsilon$, and $\mathbf{Y}$:
1. $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$.
2. $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \widehat{Y}_i$
3. $\sum_{i=1}^{n} x_i \hat{\varepsilon}_i = 0$.
4. $\sum_{i=1}^{n} \widehat{Y}_i \hat{\varepsilon}_i = 0$.
5. The regression line always goes through the point $(\bar{x}, \bar{Y})$.

The solutions, $\boldsymbol{\beta}$, to (12.5) are generally easier to express in matrix notation than in summation notation. The normal equations are now presented in matrix form. Recall that

$$\mathcal{Q} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

This is simplified first and then differentiated with respect to $\boldsymbol{\beta}$. Then, the result is set equal to $\mathbf{0}$ to solve for $\hat{\boldsymbol{\beta}}$:

$$\mathcal{Q} = \mathbf{Y}'\mathbf{Y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

Since $\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y}$ is a scalar $(1 \times 1), (\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y})' = \mathbf{Y}'\mathbf{X}\boldsymbol{\beta}$, so $\mathcal{Q}$ simplifies to

$$\mathcal{Q} = \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

The expression for $\frac{\delta \mathcal{Q}}{\delta \boldsymbol{\beta}}$ can now be calculated:

$$\begin{aligned}\frac{\delta \mathcal{Q}}{\delta \boldsymbol{\beta}} &= \frac{\delta}{\delta \boldsymbol{\beta}}\left(\mathbf{Y}'\mathbf{Y}\right) - \frac{\delta}{\delta \boldsymbol{\beta}}\left(2\left(\mathbf{X}'\mathbf{Y}\right)'\boldsymbol{\beta}\right) - \frac{\delta}{\delta \boldsymbol{\beta}}\left(\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\right) \\ &= 0 - 2\mathbf{X}'\mathbf{Y} - \left[\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)'\boldsymbol{\beta}\right]\end{aligned}$$

$= -2\mathbf{X}'\mathbf{Y} - 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$

Setting equal to zero and solving for $\boldsymbol{\beta}$ yields

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y},$$

The likelihood function for $\boldsymbol{\beta}$ and σ^2 when $\mathbf{X}$ is given is

$$\mathcal{L}\left(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-\left(Y_i - \left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1}x_{i,p-1}\right)\right)^2}{2\sigma^2}\right]$$

In matrix form, this is

$$\mathcal{L}\left(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right]$$

The natural log of the matrix form of the likelihood function (log-likelihood function) is

$$\ln \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}\right) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\sigma^2\right) - \frac{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}$$

Simplifying the partial derivative of the log-likelihood function with respect to $\boldsymbol{\beta}$ gives

$$\begin{aligned}\frac{\delta \ln \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}\right)}{\delta \boldsymbol{\beta}} &= \frac{\delta}{\delta \boldsymbol{\beta}}\left[-\frac{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right] \\ &= \frac{\delta}{\delta \boldsymbol{\beta}}\left[\frac{-\mathbf{Y}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{2\sigma^2}\right] \\ &\quad \text{Recall that } \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} \text{ is } 1 \times 1 \\ &= \frac{\delta}{\delta \boldsymbol{\beta}}\left[\frac{-\mathbf{Y}'\mathbf{Y} + 2\left(\mathbf{X}'\mathbf{Y}\right)'\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{2\sigma^2}\right]\end{aligned}$$

By Rules of Differentiation 1 and 3 on page 671

$$\begin{aligned}&= \frac{2\mathbf{X}'\mathbf{Y}}{2\sigma^2} - \left[\frac{\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)'\boldsymbol{\beta}}{2\sigma^2}\right] \\ &= \frac{\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{\sigma^2}\end{aligned}$$

Setting this equal to zero and solving for $\boldsymbol{\beta}$ yields

$$\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

The likelihood function for $\boldsymbol{\beta}$ and σ^2 when $\mathbf{X}$ is given is

$$\mathcal{L}\left(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-\left(Y_i - \left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1}\right)\right)^2}{2\sigma^2}\right]$$

In matrix form, this is

$$\mathcal{L}\left(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right]$$

The natural log of the matrix form of the likelihood function (log-likelihood function) is

$$\ln \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}\right) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\sigma^2\right) - \frac{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}$$

Simplifying the partial derivative of the log-likelihood function with respect to $\boldsymbol{\beta}$ gives

$$\begin{aligned}
\frac{\delta \ln \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}\right)}{\delta \boldsymbol{\beta}} &= \frac{\delta}{\delta \boldsymbol{\beta}}\left[-\frac{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right] \\
&= \frac{\delta}{\delta \boldsymbol{\beta}}\left[\frac{-\mathbf{Y}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{2\sigma^2}\right] \\
&\text{Recall that } \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} \text{ is } 1 \times 1 \\
&= \frac{\delta}{\delta \boldsymbol{\beta}}\left[\frac{-\mathbf{Y}'\mathbf{Y} + 2\left(\mathbf{X}'\mathbf{Y}\right)'\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{2\sigma^2}\right]
\end{aligned}$$

$$\begin{aligned}
\frac{\delta \ln \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}\right)}{\delta \boldsymbol{\beta}} &= \frac{\delta}{\delta \boldsymbol{\beta}}\left[-\frac{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right] \\
&= \frac{\delta}{\delta \boldsymbol{\beta}}\left[\frac{-\mathbf{Y}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}}{2\sigma^2}\right] \\
&\text{Recall that } \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} \text{ is } 1 \times 1 \\
&= \frac{\delta}{\delta \boldsymbol{\beta}}\left[\frac{-\mathbf{Y}'\mathbf{Y} + 2\left(\mathbf{X}'\mathbf{Y}\right)'\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{2\sigma^2}\right] \\
&\text{By Rules of Differentiation 1 and 3 on page 671} \\
&= \frac{2\mathbf{X}'\mathbf{Y}}{2\sigma^2} - \left[\frac{\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)'\boldsymbol{\beta}}{2\sigma^2}\right] \\
&= \frac{\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{\sigma^2} \\
&= \frac{\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{\sigma^2}
\end{aligned}$$

Setting this equal to zero and solving for $\boldsymbol{\beta}$ yields

$$\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}.$$

It is also of interest to find the MLE for σ^2. Taking the partial derivative of the log-likelihood function in terms of σ^2 gives

$$\frac{\delta \ln \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}\right)}{\delta \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2} \cdot (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

When this quantity is set equal to zero and solved for σ^2, the MLE is

$$\tilde{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n} = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}$$

Unfortunately, $\tilde{\sigma}^2$ is a biased estimator of σ^2. The bias is easily fixed and the unbiased estimator $\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-p}$ is typically used to estimate σ^2.

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right).$$

In Example 12.5 on page 574 , the variance of $\hat{\boldsymbol{\beta}}$ was shown to equal $\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$. Next, $\hat{\boldsymbol{\beta}}$ is shown to be an unbiased estimator of $\boldsymbol{\beta}$. Specifically,

$$\begin{aligned}
\text{If } \hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} \\
\text{Then } E[\hat{\boldsymbol{\beta}}] &= E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}\right] \\
&= E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})\right] \\
&= E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\right] \\
&= E\left[\mathbf{I}\boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon\right] \\
&= \boldsymbol{\beta} \text{ since } \mathbf{I} \text{ and } \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' \text{ are constants and } E(\boldsymbol{\varepsilon}) = \mathbf{0}.
\end{aligned}$$

under the normal error regression model. However, unbiasedness does not guarantee uniqueness. Fortunately, the Gauss-Markov theorem guarantees that among the class of linear unbiased estimators for $\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}$ is the best in the sense that the variances of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are minimized. Consequently, $\hat{\boldsymbol{\beta}}$ is called a best linear unbiased estimator, or a BLUE. Note that the error variance σ^2 is unknown, but its unbiased estimate is given above,

$$\underbrace{Y_i - \bar{Y}}_{\text{Total Deviation}} = \underbrace{\underbrace{\widehat{Y}_i - \bar{Y}}_{\substack{\text{Deviation of Fitted} \\ \text{Regression Value} \\ \text{around the Mean}}} + \underbrace{Y_i - \widehat{Y}_i}_{\substack{\text{Deviation around the} \\ \text{Fitted Regression Line}}}}$$

$2\sum_{i=1}^{n}\left(\widehat{Y}_i - \bar{Y}\right)\left(Y_i - \widehat{Y}_i\right) = 2\sum_{i=1}^{n}\left(\widehat{Y}_i - \bar{Y}\right)\hat{\varepsilon}_i = 0$

$\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left(\widehat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$

SST: Sum of squares total SSR: Sum of squares regression SSE: Sum of sqaures residue

Goodness of fit : R^2, percentage of variance explained, 1-SSR/SST, when SSR goes up it is bad prediction.

X is not full rank: cilumns are linearly dependent.

Orthogonality: X partition into X1 and X2, X1ˆTX2 = 0.

Largemodel L, Smallmodel S. RSS_S - RSS_L is small. H0: S, H1: L. maxLikeihood(L)/maxLikeihood(S), if ratio goes up, accept H1, reject H0.

dim(L) = p, dim(S) = q. F = [RSS_S-RSS_L/p-q]/[RSS_L/n-p] = degree of freedom, F_(p-q, n-p), TSS = RSS_S, q = 1.

Step 1: Hypotheses $-H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.
Step 2: Test Statistic $-\hat{\beta}_1 = 0.0030943$ is the test statistic. Assuming the assumptions of Model (12.4) are satistfied,

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2_{\hat{\beta}_1}\right)$$

The standardized test statistic under the assumption that H_0 is true and its distribution are

$$\frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \sim t_{200-2}$$

Step 3: Rejection Region Calculations - Because the standardized test statistic is distributed t_{198} and H_1 is a two-sided hypothesis, the rejection region is $|t_{\text{obs}}| > t_{0.95;198} = 1.6526$. The value of the standardized test statistic is $t_{\text{obs}} = \frac{0.0031-0}{.00019} = 15.912$.
Step 4: Statistical Conclusion - The $\wp$-value is $2 \times \mathbb{P}(t_{198} \geq 15.912) = 2 \times 0 = 0$. I. From the rejection region, reject H_0 because |15.912| is greater than 1.6526.
II. From the $\wp$-value, reject H_0 because the $\wp$-value = 0 is less than 0.10. Step 5: English Conclusion - There is evidence to suggest a linear relationship between sat and gpa.

Bootstrap: samples with replacement. no assumptions CI relies normality assumptions. Sample from obs data not true model.

PI > CI, population parameter is constant and we do not know distribution.

Autoregression: basic technique for time series

Generalized least squares (gls): errors are dependent Weighted least squares (wls): errors are ind but not iid. errors are not normal distribute: Robust regression

gls: we assumed var(E) = sigmaˆ2I, sigma unknown but sumation known. Cholesky decomposition: Var(beta) = ($X^{\text{Tsumation}}$-1X)ˆ-1 sigmaˆ2, error depend on each other due to sumation

wls is speacial case of gls: errors uncorrelated but unequal variance, let w = diagonal matrix with w on diagonal. beta = ($X^{\text{TwX)}}$-1XˆTwY

Robust regression, M-estimation, choose beta to minimize: least absolute deviation (LAD), sumation p(Y - XˆTBeta), p = xˆ2, p = |x|, Huber's method: est of sigma, w = 1/2 xˆ2 if |x| <= c, c|x| - 1/2 cˆ2 otherwise.

Least Trimmed Square (LTS): if errors large or extreme, huber fail, LTS minimize sum of squares of q smallest residuals. q -> small(1/2 n) + small(1/2(p+1)).

Broken stick regression: savings data, dramatic difference when pop 15 <= 35 : pop15 > 35 basis function: beta = c -x if x <= c, 0 otherwise. hockey sticks when guaranted to meet c

Test based procedures for model selection. Backward Elimination: full model, delete highest p-value predictor if p > avafar critical : != 5%, >= 15%, re-fit model: repeat (1 predictor at a time) Forward selection: start null model, add 1 predictor repeat for all, smallest p-value, p < avafar critical Criterion procedure: pick g() that is close to f, distance between them: kullback-liebler distance, $D_{\mathrm{KL}}(f\|g) = \int_{-\infty}^{\infty} f(x)\log\left(\frac{f(x)}{g(x)}\right)dx$ Akaike: AIC: -2l()+2p, linear regression: -2l()=nlog(RSS/n) + c?

Adjusted R^2, add variable, incentive: pick largest model, full model is the downside. R^2 = 1 - RSS/TSS, so when rss goes down, r^2 goes up. R^2 adjust = 1 - (n-1/n-p)(1-r^2) Adding a predictor will only increase R^2 adjust if the predictor has value

Dimensionality reduction PCA. Take X, delete columns of 1's, each column gets mean zero (transformation). Try to find orthogonal space: 1. find u1 such that var(u1^TX) is maximum , subject to u1^Tu1 = 1. 2. find u2 ... , u1^Tu2 = 0. Z = u^TX, principal components, rotation matrix is u^T.

Ridge regression: assume beta is small, large number of predictors. Procedure: predictors - centered, scaled by s.d.
target - centered choose beta that minimizes: $\widehat{\beta}_{\text{ridge}} = \left(X^T X + kI_p\right)^{-1} X^T y$ k creates a RIDGE down X^TX, choose beta to minimize (y-Xbeta)^T(y-xbeta) subject to sum beta^2 <= t^2

Lasso: same concept as redge except we want to choose beta to minimize:(y-Xbeta)^T(y-xbeta) subject to sum |beta| <= t

prediction and residual being orthagothic, because they should be uncorrelation.